

Weekly Report

2012.11.5-2012.11.11

黄芯芯

本周工作：

1. 完成了操作系统课程报告和数据挖掘课程报告。
2. 淘宝标签项目：

之前对 MDS 投影后的点是用户交互圈选进行聚类，即用户框选某些点使它们成为一个聚类。这周尝试对 MDS 投影后的点进行 k-means 聚类（做了之后发现，其实这是个错误的尝试）。数据点进行 MDS 投影后，用户可以根据点在二维空间上的分布选择每个聚类的中心点，然后使用 k-means 方法进行聚类，最后对聚类结果进行简单的可视化显示以查看每个聚类的特征。具体流程如下所示：

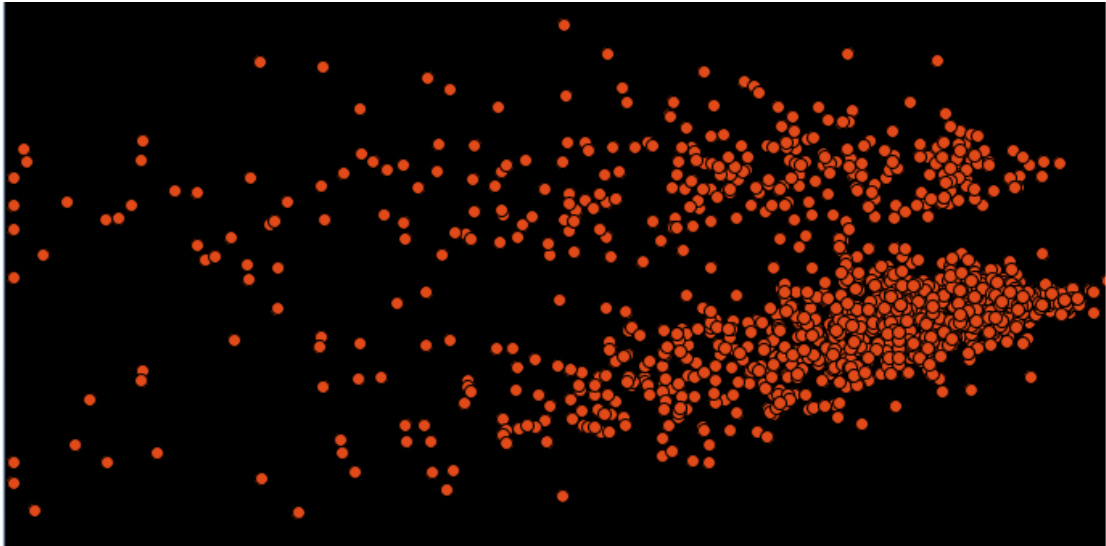


图 1 MDS 投影结果

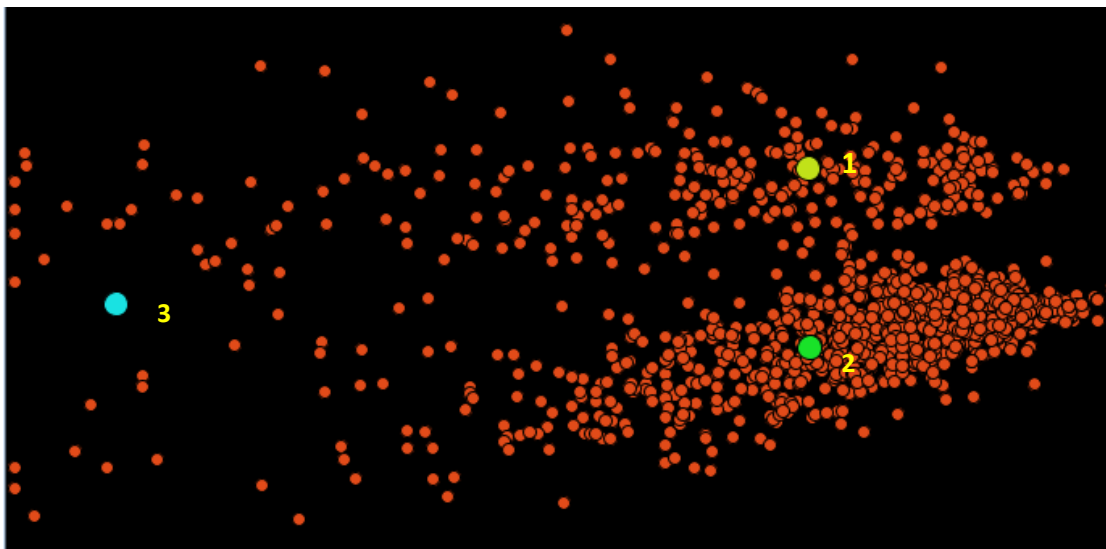


图 2 用户指定 3 个种子点

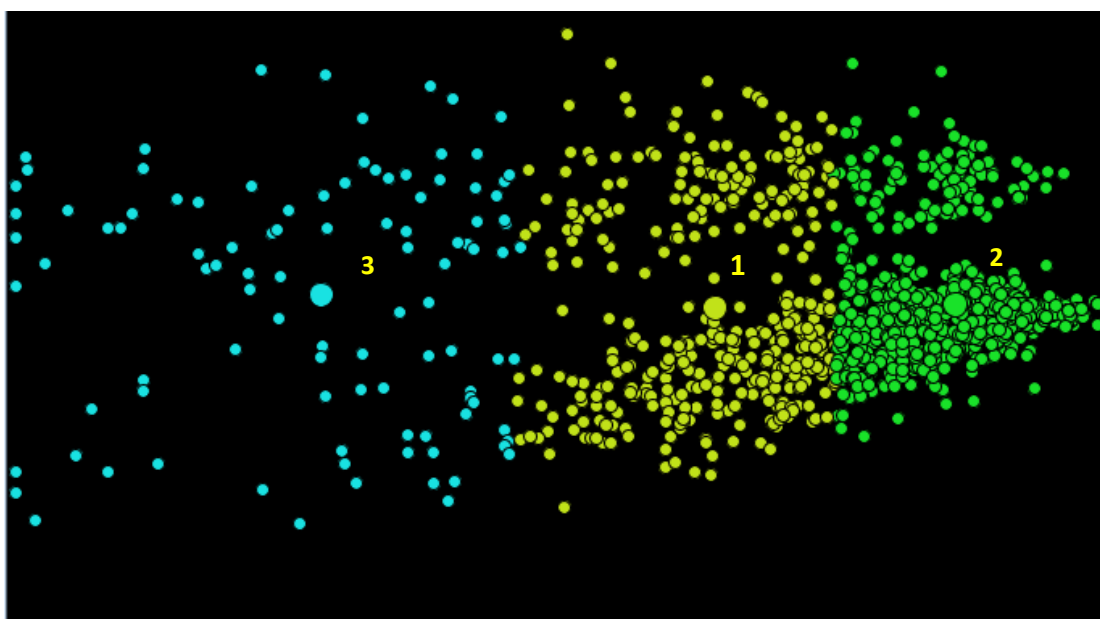


图 3 k-means 聚类结果

原本想要的聚类结果是右边上下各为一个聚类，然后左边分散比较零散的点成为一个聚类。但是从图 3 使用 k-means 聚类的结果来看，并没有得到想要的结果。

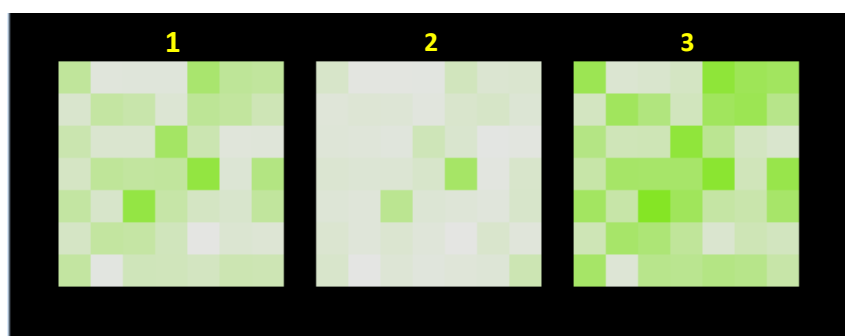


图 4 三个聚类的可视化结果



图 5 第二个聚类中购买比率较高的类目

从聚类的结果上看，其实只有第二个聚类的特征比较明显，这个聚类中的用户除了“移动/联通/充值”和“3C 数码配件市场”两个类目购买比率较高之外，其他类目的购买比率都非

常低。第三个聚类的点本来分布就比较零散，所以从可视化的结果上看也几乎没有特别的购买特征，购买的类目覆盖率比较高。

选取四个初始质心点，将数据点分成四个聚类，得到如下的结果：

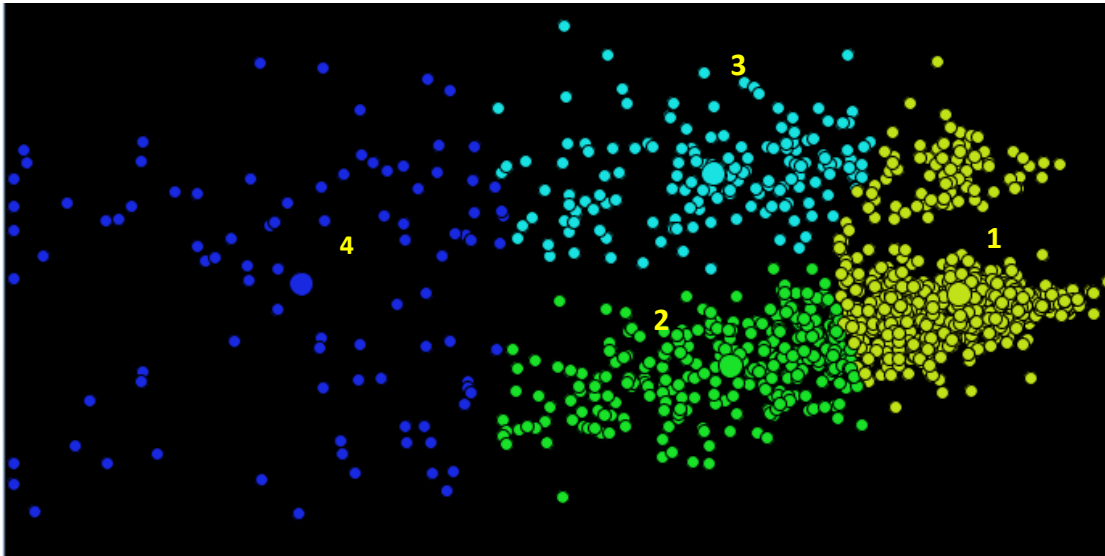


图 6 当 $k=4$ 时的聚类结果

将聚类结果进行可视化如下图所示：

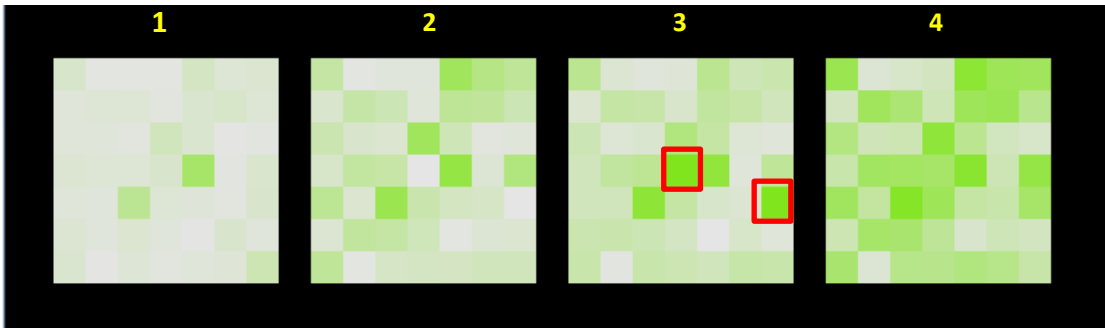


图 7 当 $k=4$ 时的聚类结果可视化

在图 7 的第三个聚类中，可以发现红框出两个类目的颜色特别深，即该聚类下的买家在这两个类目下的购买比率比较高。查看这两个类目，如下图 8 所示：

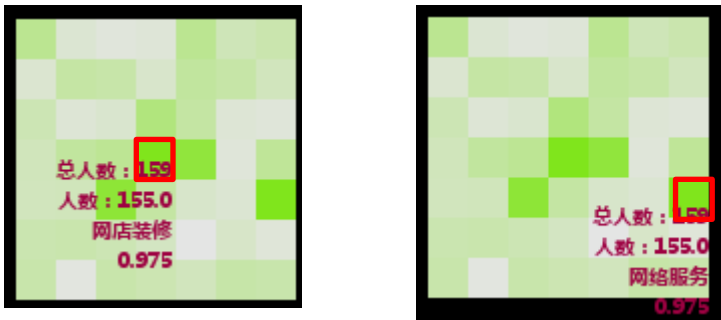


图 8 第三个聚类中购买比率较高的类目

发现这两个类目分别是“网店装修”和“网络服务”，而且购买占比非常高，整个聚类的人数为 159 人，而购买这两个类目的人数达到 155 人。因此，有理由推测这个聚类下的这些用户其实也是淘宝的卖家。

以上是这周对 MDS 投影的结果进行 k-means 的尝试，结果还不如之前用户直接圈选的结果好。从原来的 MDS 投影结果来看，右边上下两个明显的聚类并没有被聚出来，问题出在 MDS 投影的结果对 k-means 方法有影响，不同的 MDS 结果会产生不同的聚类。但是对与 k-means 方法需要预先设定 k 值的缺点来说，进行 MDS 之后可以是用户在二维空间上直观地看到整个数据集大约应该分成多少类，即可以帮助用户决定 k 的值，当 k 的值设定后再采用 k-means 方法在全部维度上进行聚类，这样得到的结果应该会比较好一点。

下周计划：

1. 淘宝那边要帮 TCIF 做一个关于用户“爱好”可视化展示，使用 d3。
2. 改进筛选树视图，加入马赛克图模块。